Lab 3 – `dplyr`

Contents


```
## Copy and paste the code chunk below into the top of your R Markdown document
library(ggplot2)
library(dplyr)

theme_set(theme_bw(base_size = 14))

## College data
college <- read.csv("https://collinn.github.io/data/college2019.csv")

## For this lab, we don't need all of the columns so we will
# select only a few of them
college <- select(college, Name, State, Enrollment, Type,
                  Region, Adm_Rate, ACT_median, Cost, Net_Tuition,
                  Avg_Fac_Salary, Debt_median)

## Load data from ggplot package
data(mpg)
```

Problem Sets

Question 1: Filter the college dataset to include only schools located in the Plains and the Great Lakes regions and with enrollments less than 20,000. Using your filtered data, create a two-way table with the variables `Region` and `Type`. Based on this table, in which region do most private schools tend to be located?

Solution:

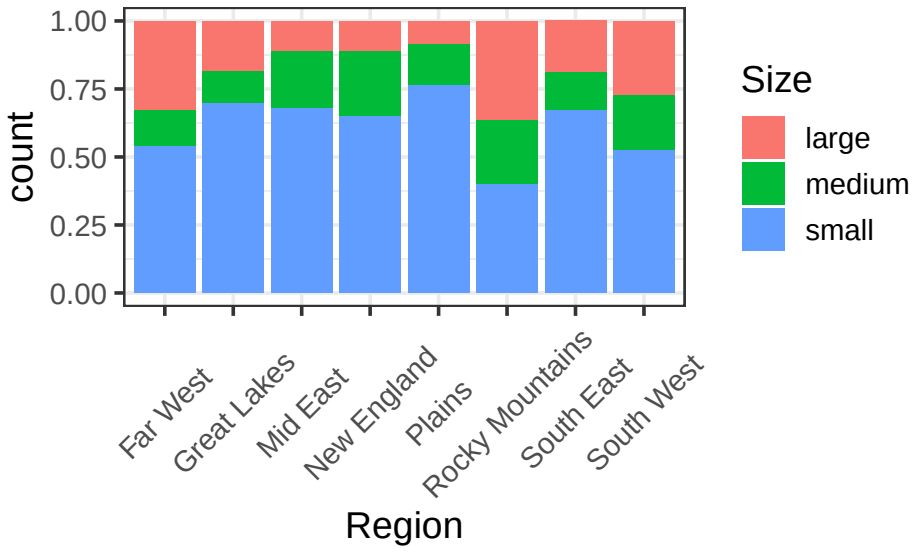
```
filter(college, Region %in% c("Plains", "Great Lakes"),
       Enrollment < 20000) %>% with(table(Region, Type))
```

```
##              Type
## Region        Private Public
##   Great Lakes     125     50
##   Plains           84     37
```

Question 2: Using the college dataset, create a new variable called `Size` that takes the values `"small"` when enrollment is less than 5000, `"medium"` when enrollment is 5,000 or greater but less than 10,000, and `"large"` otherwise. Then create an appropriate bar chart to determine which region has the greatest proportion of small schools.

Solution:

```
tt <- college %>%
  mutate(Size = case_when(
    Enrollment < 5000 ~ "small",
    Enrollment >= 5000 & Enrollment < 10000 ~ "medium",
    Enrollment >= 10000 ~ "large"
  ))
ggplot(tt, aes(fill = Size, Region)) + geom_bar(position = "fill") +
  theme(axis.text.x = element_text(vjust = 0.5, angle = 45))
```



Question 3: Using the tips dataset, do the following:

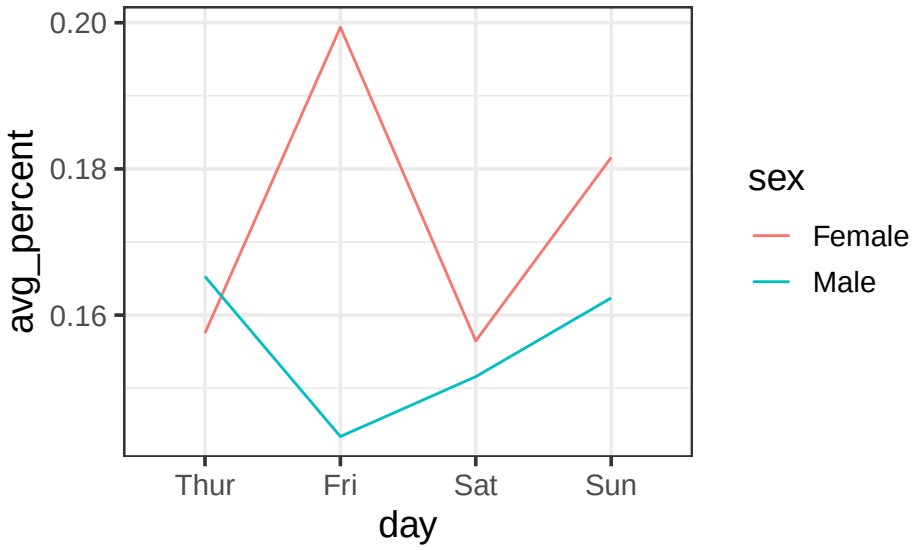
- Create a new variable, `percent_tip` that computes the percentage tip for each meal
- Summarize the dataset to show the average tip percent by sex and day of week
- Create a line chart using `geom_line()` to show trends over time
  - Hint: Use both the `color` and `group` aesthetic in `aes()`
  - Hint: Use `scale_x_discrete()` so that the days on are increasing order (Thur, Fri, Sat, Sun)
- Based on the graphic you created, what are some observations that you make relating tip percentage to sex and day of week?

```
tips <- read.csv("https://collinn.github.io/data/tips.csv")
```

Solution:

```
tt <- mutate(tips, percent = tip / total_bill) %>%
  group_by(sex, day) %>%
  summarize(avg_percent = mean(percent))

ggplot(tt, aes(x = day, y = avg_percent, color = sex, group = sex)) +
  geom_line() +
  scale_x_discrete(limits = c("Thur", "Fri", "Sat", "Sun"))
```



Question 4: Intensive care units, or ICUs, are primary spaces in hospitals that are reserved for patients in critical condition. The data below is a random sample of n = 200 ICU patients from a research hospital affiliated with Carnegie Mellon University (CMU).

```
icu <- read.csv("https://collinn.github.io/data/icuadmit.csv")
```

Descriptions of the relevant variables are indicated below:

- **ID** - Patient ID number
- **Status** - Patient status: 0=lived or 1=died
- **Age** - Patient's age (in years)
- **Infection** - Is infection involved? 0=no or 1=yes
- **Previous** - Previous admission to ICU within 6 months? 0=no or 1=yes

Using the `icu` dataset and the functions described in this lab, complete the following stages:

1. Change the `Previous` variable to have values `"no"` and `"yes"` instead of `0` and `1`
2. Filter the data to *only* include patients whose visit involves an infection
3. For the `Age` variable, find the mean, standard deviation, and group size (found using the function `n()`) of patients with and without a previous admission to the ICU in the prior 6 months.

Your solution should indicate the total number of patients with and without previous admission, along with each group's mean age and standard deviation. It will contain two rows and four columns – your final output should look like this:

Solution:

```
icu %>% filter(Infection == 1) %>%
  mutate(Previous = ifelse(Previous == 0, "no", "yes")) %>%
  group_by(Previous)  %>%
  summarize(MeanAge = mean(Age),
            SDAge = sd(Age),
            N = n()) %>% as.data.frame()
```

```
##   Previous MeanAge  SDAge  N
## 1       no  62.369 17.842 65
## 2      yes  57.000 17.601 19
```